Séparation texte/graphique à partir d'une représentation parcimonieuse

Thai V. Hoang, Salvatore Tabbone

Séparation texte/graphique à partir d'une représentation parcimonieuse

Thai V. Hoang[*,**] — **Salvatore Tabbone**[**]

[*] *Institute Polytechnique de Hanoi*
MICA, UMI 2954, Hanoi, Vietnam

[**] *Université Nancy 2, LORIA, UMR 7503*
54506, Vandœuvre-lès-Nancy, France
{vanthai.hoang,tabbone}@loria.fr

RÉSUMÉ. *Nous proposons dans ce papier une méthode pour séparer le texte du graphique dans des documents techniques. Nous considérons que le texte et le graphique représentent un signal bidimensionnel pour lequel chaque composante a des caractéristiques différentes. L'algorithme que nous proposons repose sur une représentation parcimonieuse du document technique pour lequel deux dictionnaires (ensemble de vecteurs surcomplet) appropriés sont définis chacun donnant une représentation parcimonieuse pour une composante du signal et pas pour l'autre. Des heuristiques sont également proposées pour regrouper le texte en chaînes de caractères. Les résultats expérimentaux obtenus sur différents types de documents sont très prometteurs par rapport à des approches purement géométriques.*

ABSTRACT. *A novel text extraction method from graphical document images is presented in this paper. Graphical document images containing text and graphics components are considered as two-dimensional signals by which text and graphics have different morphological characteristics. The proposed algorithm relies upon a sparse representation framework with two appropriately chosen discriminative overcomplete dictionaries, each gives sparse representation over one type of signal and non-sparse representation over the other. Separation of text and graphics components is obtained by promoting sparse representations of input images in these two dictionaries. Some heuristic rules are used for grouping text components into text strings in post-processing steps. The proposed method overcomes the problem of touching between text and graphics. Experiments show some promising results on different types of documents.*

MOTS-CLÉS : *Séparation texte/graphique, transformée en ondelettes, représentation parcimonieuse*

KEYWORDS: *Text/graphic separation, wavelet transform, sparse representation*

1. Introduction

La séparation texte/graphique constitue une étape majeure dans le traitement des documents techniques. Il s'agit de les séparer en deux couches : le graphique d'une part et le texte d'autre part. Extraire le texte est essentiel car le texte véhicule des concepts sémantiques sur le document qui peuvent être exploités par un humain ou mis en évidence par un OCR. Dans ce contexte, il est important de mettre en œuvre des approches robustes qui permettent de traiter automatiquement tout types de documents techniques tels que les plans architecturaux, électriques, les chèques postaux,...

Souvent ces documents contiennent au moins les deux couches texte et graphique. La couche graphique contient toute la symbolique qui peut-être plus ou moins complexe suivant l'application et cette couche se décompose elle-même en atomes comme des lignes, des courbes, des polygones, des cercles... Le texte quant à lui est composé de chiffres et caractères regroupés en chaînes de caractères utilisés pour annoter le graphique. Extraire le texte est un challenge pour différentes raisons :

- les atomes peuvent être de taille, d'orientation et d'épaisseur quelconques. Le texte peut également être dans une fonte et orientation quelconque et différente au sein d'un même document,

- le texte peut toucher la couche graphique. Certains types de documents sont intrinsèquement denses et les caractères se superposent aux autres couches et sont parfois occultés.

De nombreuses méthodes ont été proposées dans cette problématique. La plus connue étant celle de Fletcher et Kasturi [FLE 88] qui se base sur des critères géométriques définis sur les composantes connexes et reposant sur un ratio lié à la taille des composantes connexes des caractères. L'idée sous-jacente principale est que la distribution de la taille des composantes connexes textes est différente de celle du graphique. L'approche de Fletcher et Kasturi a montré son efficacité et sa robustesse pour des documents avec des caractères de tailles différentes et d'espacements différents. Cependant, lorsque les caractères collent au graphique les approches qui s'appuient sur la taille des composantes connexes sont inefficaces et il est nécessaire de mettre en œuvre des post-traitements. Pour ce faire, certaines approches [GLO 92, LI 00, VEL 03] font évoluer en taille le rectangle englobant des caractères trouvés dans la couche texte pour récupérer au voisinage de ces derniers les caractères qui collent au graphique. Ensuite le caractère est décroché soit en coupant aux frontières du rectangle englobant [LI 00] soit en se basant sur la ligne de portée des chaînes de caractères [VEL 03] ou sur l'alignement des caractères à partir de la transformée de Hough [GLO 92]. Des améliorations peuvent être apportées sur la recherche de voisinage en tenant compte du diagramme de voisinage de Voronoï [WAN 01] ou sur le point de décrochage du caractère à partir du squelette du document [CAO 01]. Dans [TOM 02] les auteurs s'appuient également sur le squelette pour décrocher le caractère mais en se basant sur des critères de Gestalt de continuité et de proximité pour identifier les caractères qui touchent le graphique. Récemment Su et al. [SU 09] s'ap-

puie sur une méthode de vectorisation [SON 02] pour identifier dans des documents techniques des lignes qui touchent les caractères.

D'autres approches originales et duales consistent à soustraire du document le graphique à partir de techniques de filtrage morphologique [LUO 97] ou géométrique des composantes du document [LU 98, SON 02]. Soulignons également une approche de suivi de courbes principales à partir d'un graphe de voisinage des pixels avec des techniques de réécriture pour simplifier le graphe [CHE 04]. Des approches multi-résolution ont également été proposées. La première par Olivier et Dominique [DEF 94] pour traiter du courrier postal et ensuite adapté par Tan et Ng [TAN 98]. L'idée est de faire ressurgir naturellement au niveau le plus haut de la pyramide les groupes de mots. Cependant quand le texte et le graphique sont proches cette approche induit de fausses détections.

Dans ce papier nous présentons un algorithme robuste pour extraire les composantes texte de documents techniques utilisant une représentation parcimonieuse. La façon d'aborder le problème est différente des approches présentées ci-dessus qui ont été mises au point pour des applications spécifiques et difficilement applicables à d'autres applications. Nous abordons le problème de manière générale et ceci nous permet de traiter des documents très denses comme ceux de la figrue 4. Ainsi, nous considérons une image de document comme un signal bidimensionnel y qui est composé d'une mixture de deux signaux de même taille : y_t contenant le texte et y_g contenant le graphique. Le problème d'extraction du texte peut maintenant être vu comme un problème multidimensionnel de séparation aveugle de sources [JUT 91].

Pour résoudre ce problème nous utilisons l'algorithme d'analyse en composante morphologique (MCA) proposé par Starck et al. [STA 05]. Le MCA permet une bonne séparation de caractéristiques contenues dans une image lorsque ces caractéristiques présentent des aspects morphologiques différents. Pour ce faire nous nous appuyons sur une représentation parcimonieuse définie à partir de deux dictionnaires appropriés et redondants uniquement pour chacune des caractéristiques.

Ensuite nous proposons un post-traitement pour regrouper le texte issu de la couche y_t en chaînes de caractères. Enfin, l'approche que nous proposons est robuste lorsque le texte touche le graphique et la composante texte peut être dans n'importe qu'elle direction, de forme et de taille quelconque.

Le reste du papier est organisé comme suit. Des concepts sur la représentation parcimonieuse de signaux sont donnés au paragraphe 2. Ensuite nous présentons l'algorithme de décomposition et le dictionnaire sélectionné (§3). Une étape de post-traitement est proposée pour convertir le texte extrait en chaîne de caractères au paragraphe 4. Ensuite, les résultats expérimentaux sont donnés (§5) et nous concluons au paragraphe 6.

2. Représentation parcimonieuse de signaux

Cette notion de parcimonie prend racine dans le système de vision humain où le cortex visuel a été caractérisé comme étant spatialement localisé, orienté et passe-bande comparable à des bases de fonction de transformée en ondelette. En plus il a une aptitude à produire une distribution parcimonieuse en réponse à des images naturelles [OLS 96]. Cette théorie a été validée par Olshausen et Field [OLS 98] qui a considéré le problème de codage efficace d'images naturelles. Les auteurs ont montré que lorsqu'un dictionnaire de code est redondant (le nombre d'élément du code est beaucoup plus important que l'espace auquel il appartient) et non-orthogonal, une stratégie de codage où il est question de maximiser l'utilisation d'un petit nombre d'éléments du dictionnaire revient à sélectionner seulement les codes nécessaires à représenter une observation donnée. Selon le principe de réduction de redondance de Barlow [BAR 89], le code parcimonieux obtenu est une représentation plus efficace pour des traitements ultérieurs.

Le principe est donc de modéliser un vecteur d'observation par un petit nombre de signaux élémentaires appartenant à un ensemble surcomplet de vecteurs. Cet ensemble est souvent appelé dictionnaire redondant au sens que nous venons de définir ci-dessus. Ainsi il existe des relations linéaires entre certains éléments de cet ensemble et il ne s'agit donc pas de décomposer le signal sur une base au sens propre du terme, mais plutôt de rechercher la représentation la plus exacte possible tout en utilisant un petit nombre d'élément du dictionnaire.

Cette idée a été formulée mathématiquement. Soit une image observée y de taille $w \times h$ et soit le vecteur $b \in \mathbb{R}^n (n = wh)$ dans lequel les colonnes de la matrice ont été empilées ; soit un dictionnaire redondant $A \in \mathbb{R}^{n \times K}$ avec $n \ll K$ permettant une représentation parcimonieuse de b et soit x la représentation de b dans A satisfaisant $b = Ax$. Ainsi définir une représentation parcimonieuse $\hat{x}$ de b dans A est équivalent à résoudre le problème d'optimisation suivant :

$$\min_x \|x\|_0 \quad \text{sous condition que} \quad Ax = b \qquad [1]$$

où $\|.\|_0$ est la norme ℓ_0 correspondant au nombre de composantes non nulle de x. Le dictionnaire redondant A qui permet une représentation parcimonieuse du signal observé b peut-être définie de différentes façons : soit à partir d'un ensemble de signaux exemples [AHA 06] soit à partir d'un ensemble de signaux prédéfini comme la transformée en ondelette, en ondelette non décimée, en curvelet, de Gabor, en Fourier, en Fourier à fenêtre glissante,...

Du fait de la norme ℓ_0 ce problème (eq. (1)) est non convexe, combinatoire et en général NP-complet [DAV 97]. Cependant des travaux récents [DON 06] ont montré qu'en remplaçant la norme ℓ_0 par la norme ℓ_1 qui est connue pour ces propriétés de parcimonie, l'équation 1 s'écrit alors :

$$\min_x \|x\|_1 \quad \text{sous condition que} \quad Ax = b. \qquad [2]$$

Si la condition $b = Ax$ est remplacé par $b = Ax + z$, où $z \in \mathbb{R}^n$ est un terme de bruit avec $\|z\|_2 < \varepsilon$, pour tenir compte de possible petites inclusion de bruit dense sur l'image d'entrée y ou de petites erreurs dans la représentation, l'equation (2) est modifiée comme suit :

$$\min_x \|x\|_1 \quad \text{sous condition que} \quad \|Ax - b\|_2 \leq \varepsilon. \qquad [3]$$

Maintenant il s'agit d'un problème convexe d'optimisation qui peut s'écrire par la formulation quadratique :

$$\min_x \frac{1}{2}\|Ax - b\|_2 + \lambda\|x\|_1, \qquad [4]$$

où le paramètre λ est un multiplicateur de Lagrange qui permet de régler le degré de parcimonie de $\hat{x}_1$ et l'erreur de représentation car ce paramètre est relié à l'erreur ε. Des algorithmes récents efficaces en temps de calcul ont été proposés pour la résolution de ce problème quadratique [CHE 98].

3. Séparation texte/graphique à partir d'une analyse en composantes morphologiques

Il s'agit d'un algorithme permettant de résoudre de façon itérative le problème d'optimisation présenté dans l'équation (8) ci-dessous dans le cadre d'une représentation parcimonieuse. Cet algorithme a montré son utilité pour décomposer des images en texture et en structure lisse par morceaux ou pour des applications de restauration[1][ELA 05, FAD 09]. En analyse de documents, cet algorithme a été utilisé par Pan et al. [PAN 07] dans le cas de document textuel pour extraire le texte sur un fond non homogène.

3.1. *Analyse en composantes morphologiques*

Soit un signal $b \in \mathbb{R}^n$ étant une combinaison linéaire de deux autres signaux b_1 et b_2 tel que $b = b_1 + b_2$ où b_1 et b_2 représentent deux types de signaux différents. Supposons qu'il existe deux dictionnaires redondants $A_1, A_2 \in \mathbb{R}^{n \times K}$ vérifiant les deux conditions :

1) Pour $i = 1, 2$

$$\min_{x_i} \|x_i\|_1 \quad \text{sous condition que} \quad A_i x_i = b_i, \qquad [5]$$

conduit à une représentation parcimonieuse $\hat{x}_i$ de b_i dans A_i.

1. Traduit de Inpainting.

2) Pour $i \neq j$

$$\min_{x_i} \|x_i\|_1 \text{ sous condition que } A_j x_i = b_i, \quad [6]$$

conduit à une représentation non-parcimonieuse $\hat{x}_i$ de b_i dans A_j.

Dans ce cas, deux dictionnaires A_1 et A_2 sont dit discriminant dans le cadre d'une représentation parcimonieuse pour des contenus de type différents. L'algorithme d'analyse en composantes morphologique permet de résoudre le problème d'optimisation suivant :

$$\min_{x_1,x_2} (\|x_1\|_0 + \|x_2\|_0) \text{ sous condition que } A_1 x_1 + A_2 x_2 = b, \quad [7]$$

qui peut être convertie en :

$$\min_{x_1,x_2} (\|x_1\|_1 + \|x_2\|_1 + \lambda \|b - A_1 x_1 - A_2 x_2\|_2) . \quad [8]$$

La solution de l'équation (8) donne $\hat{x}_1$ et $\hat{x}_2$ la représentation parcimonieuse de b_1 et b_2 respectivement dans A_1 et A_2. Ceci signifie aussi que le signal original b a été séparé en deux parties $A_1\hat{x}_1$ et $A_2\hat{x}_2$ qui correspondent à des approximations de b_1 et b_2 respectivement. L'algorithme du MCA est inspiré du Block-coordinate relaxation (BCR) de Sardy et al. [SAR 00] qui propose une méthode numérique rapide et qui requiert uniquement des multiplications de matrices-vecteurs avec des transformations unitaires et leurs inverses. L'approche de BCR se base sur la méthode de rétraction de coefficients d'ondelette de Donoho et Johnstone [DON 94].

3.2. *Sélection des dictionnaires*

Le succès de l'algorithme MCA est garanti si les deux conditions de minimisation établies en équations (5) et (6) sont respectées. Ainsi, la sélection de deux dictionnaires appropriés est essentielle. Pour des raisons de complexité numérique, A_1 et A_2 doivent posséder une calculatoire rapide. En nous basant sur les travaux antérieurs sur la parcimonie, il nous semble judicieux de choisir la transformée en curvelet [CAN 02] comme dictionnaire adapté au graphique et reconnu pour capturer efficacement les contours d'une image et la transformée en ondelette non décimée pour le dictionnaire dédié au texte et qu'on pourrait assimiler à des textures locales isotrope.

3.2.1. *Transformée en ondelette non décimée (TOD)*

Il s'agit d'une version non décimée de la transformée orthogonale en ondelette (TOO) obtenu en ne tenant pas compte du pas de décimation dans la décomposition. La TOD permet de compenser le manque d'invariance à la translation de la TOO. Contrairement à la TOO, la transformée TOD peut être représentée par une matrice avec plus de colonnes que de lignes. Le facteur de redondance (i.e. le rapport entre le nombre de colonnes sur le nombre de lignes) est $3J + 1$ où J est le nombre d'échelles.

La TOD est censée donner une représentation parcimonieuse pour des caractéristiques isotrope et non-parcimonieuse pour des caractéristiques fortement anisotrope. L'algorithme à trous de Shensa [SHE 92] donne un moyen efficace pour implémenter la TOD.

3.2.2. *Transformée en curvelet*

Soient deux fonctions continues, non-négative et réel $W(r)$ et $V(t)$ définies sur deux compacts $[1/2, 2]$ et $[-1, 1]$ respectivement vérifiant les conditions d'admissibilité avec $r > 0$ et $t \in \mathbb{R}$:

$$\sum_{j=-\infty}^{\infty} W^2\left(2^j r\right) = 1, \sum_{l=-\infty}^{\infty} V^2\left(t - l\right) = 1. \quad [9]$$

A chaque échelle j, la curvelet mère φ_j est définie par :

$$\hat{\varphi}_j(r, \theta) = 2^{-3j/4} W\left(2^{-j} r\right) V\left(\frac{2^{\lfloor j/2 \rfloor}\theta}{2\pi}\right). \quad [10]$$

Une curvelet à l'échelle j, l'orientation θ_l et la position $\mathbf{x}_{\mathbf{k}}^{j,l} = R_{\theta_l}^{-1}\left(2^{-j}k_1, 2^{-j/2}k_2\right)$ est définie par :

$$\varphi_{j,l,\mathbf{k}}(\mathbf{x}) = \varphi_j\left(R_{\theta_l}\left(\mathbf{x} - \mathbf{x}_{\mathbf{k}}^{j,l}\right)\right), \quad [11]$$

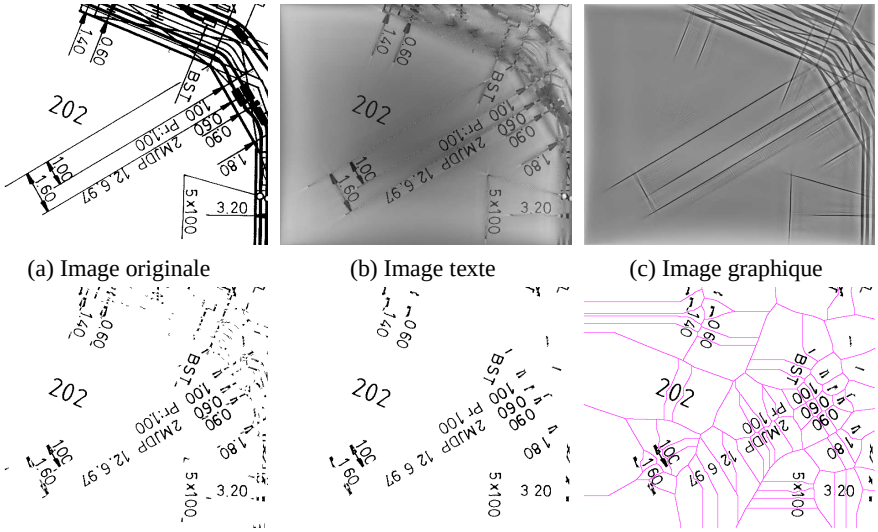
où R_{θ_l} est l'opérateur de rotation par $\theta_l = 2\pi 2^{\lfloor j/2 \rfloor} l$ radians avec $l \in \mathbb{Z}^+$ de telle sorte que $0 \leq l < 2\pi$. Les coefficients curvelet correspondant de $f \in L^2\left(\mathbb{R}^2\right)$ sont définis par le produit scalaire :

$$c_{j,l,\mathbf{k}} = \langle f, \varphi_{j,l,\mathbf{k}} \rangle = \int_{\mathbb{R}^2} f(\mathbf{x}) \overline{\varphi_{j,l,\mathbf{k}}(\mathbf{x})} d\mathbf{x}. \quad [12]$$

Il s'agit d'une transformation mulit-échelle, multi-directionnelle et allongée, obéissant à la relation d'échelle parabolique ($largeur = longueur^2$) et présentant un comportement oscillant dans la direction perpendiculaire à leur orientation. Une structure en curvelet peut être utilisée comme un dictionnaire redondant avec un facteur de redondance $16J + 1$ où J est le nombre d'échelles. Elle est censée donner une représentation parcimonieuse pour des caractéristiques anisotropes et des morceaux lisses de courbes et de droites de longueurs différentes.

3.3. *Extraction de la couche contenant le texte*

Supposons que le document initial y est décomposé en deux images de même taille : y_t pour la composante texte et y_g pour la composante graphique. L'application de l'algorithme MCA sur y utilisant la transformée non décimée et en curvelet



(a) Image originale (b) Image texte (c) Image graphique

(d) Image binaire texte (e) Image binaire texte sans les petites composantes (f) Diagramme de Voronoï en régions

Figure 1. *Étapes de traitement pour extraire des chaînes de texte de l'image*

comme dictionnaires surcomplet donnera $\tilde{y}_t$ et $\tilde{y}_g$ comme approximation de y_t et y_g. Par exemple, si on suppose que y correspond au document technique de la figure 1a la décomposition donnera respectivement pour $\tilde{y}_t$ et $\tilde{y}_g$ les images de la figure 1b et 1c. Nous pouvons constater que les deux composantes ne sont pas totalement séparées pour deux raisons :

- il y a un recouvrement entre les deux dictionnaires. Les deux considèrent les composantes basse fréquence du document qui se retrouvent aussi bien dans la couche texte que graphique,

- certains éléments graphiques (flèches, petits segments) possèdent des caractéristiques similaires aux composantes texte. Ces derniers peuvent apparaître dans l'image de texte.

Pour lever une partie de ces ambiguïtés (i.e. recouvrement entre les deux dictionnaires et similarité entre caractéristiques), nous proposons, dans la section suivante, un post-traitement qui consiste à regrouper les caractères extraits en chaînes de caractères.

4. Regroupement des caractères extraits en chaînes de caractères

Soit l'image de texte de la figure 1b issue de la méthode présentée au paragraphe 3, cette figure est convertie en binaire dans la figure 1d par seuillage adaptative

[GON 01]. La figure 1e est obtenue en éliminant les petites composantes connexes à partir de la figure 1d. La taille de la fenêtre de seuillage adaptative et la taille des petites composantes à élimiminer sont sélectionnées expérimentalement.

L'objectif est de regrouper les caractères et d'automatiser leur interprétation par un OCR dédié. L'intérêt est aussi de récupérer dans cette couche les petites composantes éliminées lors du seuillage faisant parties de la couche texte comme le '.' et le ':' qui appartiennent à la cotation et également d'éliminer le bruit (i.e. les petites composantes qui sont dans la couche texte mais qui ne correspondent à rien). Pour ce faire, nous proposons des heuristiques qui s'appuient sur le voisinage, la distance inter-caractères, l'orientation et le recouvrement. Nous partons de l'hypothèse que les caractères de la couche texte sont linéairement alignés dans les documents techniques.

Voisinage : Nous supposons qu'une chaîne de caractères contient des caractères qui sont localement voisins. Le voisinage est déterminé à partir du diagramme de Voronoï [HOA 09] de l'image binaire de la figure 1e et illustré sur la figure 1f. Par définition, chaque composante connexe est délimitée par une région de Voronoï pour laquelle tous les points sont les plus proches de cette composante connexe que des autres. Ainsi deux composantes textes sont dites voisines si elles ont des régions adjacentes.

Distance inter-caractères : La distance inter-caractères $d(g_i, g_j)$ entre deux composantes textes g_i, g_j dans une chaînes de caractères est définie par :

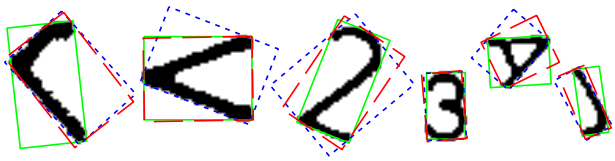
$$d(g_i, g_j) = \min_{p \in g_i, q \in g_j} d(p, q) \qquad [13]$$

Pour éviter des associations de caractères parfaitement alignés mais éloignés, cette distance dépend de la hauteur des deux composantes $h(g_i)$ et $h(g_j)$ telle que $d(g_i, g_j) < T_d \max\{h(g_i), h(g_j)\}$. La valeur du seuil T_d est déterminée expérimentalement et fixée à 1.2 dans notre cas.

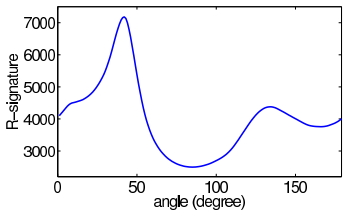
Orientation : Les composantes textes appartenant à la même chaîne de caractères doivent avoir la même orientation. Comme il n'y a pas de méthode universelle pour déterminer l'orientation d'une composante connexe nous proposons de combiner deux approches : le rectangle minimum englobant (RME) [FRE 75] et la R-signature [TAB 06]. Les RME sont présentés sur la figure 2a en couleur verte. Nous constatons que pour certains caractères l'orientation de ce rectangle est bonne par rapport à l'orientation de la composante connexe mais l'approche échoue pour des caractères telles que le 'A', 'r', 'J', etc.

Pour ce genre de caractères qui ont un trait dominant leur orientation est déterminée par la R-signature qui donne un maximum dominant dans la signature. Par exemple la figure 2b donne la R-signature du caractère 'r' et la position du maximum dans la R-signature indique l'angle qui correspond à l'orientation du 'r'. Sur la figure 2a les rectangles englobants obtenus à partir de la R-signature sont dessinés en bleu.

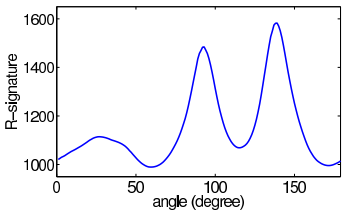
Les caractères symétriques tels que le 'A', le 'x' et le 'V' ont une R-signatures symétrique. Leurs orientations sont déterminées par l'angle qui coupe la R-signature en deux signatures de même longueur et pour lesquelles leur corrélation est la plus élevée.



(a) Rectangles englobants



(b) R-signature du caractère ‘r’



(c) R-signature du caractère ‘A’

Figure 2. *Détermination de l’orientation*

La figure 2c montre la R-signature pour le caractère ‘A’ et les rectangles englobants déterminés par cette orientation sont présentés sur la figure 2a en rouge. Désormais soient $[o_{i1}, o_{i2}, o_{i3}]$ les trois orientations déterminées par les trois méthodes énoncées ci-dessus. La différence d’orientation entre deux composantes gi,gj est définie par :

$$O_{ij} = \min_{1 \leq m,n \leq 3} |o_{im} - o_{jn}|. \quad [14]$$

Ainsi deux composantes texte g_i et g_j doivent satisfaire la condition $O_{ij} \leq T_o$ pour prétendre appartenir à la même chaîne de caractères. La valeur de T_o est déterminée expérimentalement et fixée à 0.15 radian.

Recouvrement : Deux composantes textes g_i, g_j d’une même chaîne de caractères sont voisines si elles se recouvrent en fonction d’une orientation commune déterminée par la bissectrice t_{ij} correspondant à l’angle formé par les deux lignes d’orientations déterminées par le rectangle englobant de g_i, g_j. Soient $[a_i, b_i]$ et $[a_j, b_j]$ les segments correspondant respectivement à la projection orthogonale de g_i et g_j sur t_{ij} (montré sur la figure 3). Le degré de recouvrement entre les deux composantes g_i et g_j est calculé ainsi :

$$L_{ij} = \frac{max\{min(b_i - a_j, b_j - a_i), 0\}}{min(b_i - a_i, b_j - a_j)}. \quad [15]$$

Le numérateur de l’équation (15) correspond à la longueur du segment de recouvrement. Ainsi deux composantes textes g_i et g_j doivent satisfaire $L_{ij} \geq T_l$ pour être considéré appartenant à la même chaîne de caractères. La valeur de T_l est déterminée expérimentalement et fixée à 0.75.

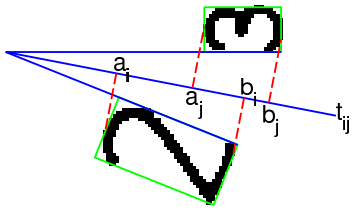


Figure 3. *Détermination de le recouvrement*

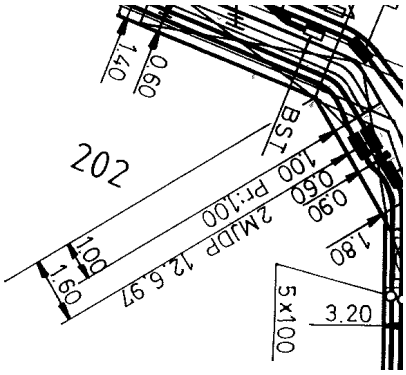
5. Résultats expérimentaux

Pour montrer la robustesse de notre approche nous l'avons évalué sur la même base d'exemples utilisés dans [TOM 02]. Cette base est composée des cinq documents techniques présentés sur la figure 4 (la première colonne). La deuxième colonne de cette figure montre le résultat final après binarisation des images textes obtenus par l'algorithme MCA et puis élimination des petites composantes. Nous pouvons constater d'après ces résultats que la séparation texte/graphique utilisant l'algorithme MCA parvient à extraire des caractères qui touchent le graphique et ceci quelques soient la fonte, le style et l'orientation des caractères.

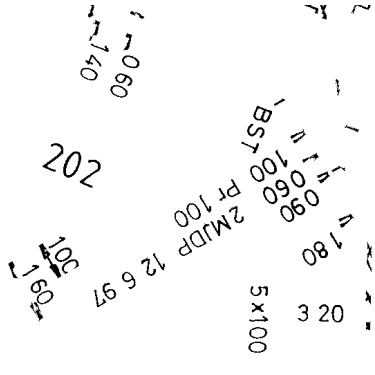
Nous avons procédé à une évaluation quantitative de notre méthode. La vérité terrain a été réalisée à la main et nous considérons les caractères, les chiffres et tous les caractères de ponctuation utiles à la cotation ('.', ':', 'x'...). La mesure que nous avons évaluée est le taux de rappel des composantes textes. Dans une optique d'interprétation/indexation automatique, nous pensons qu'il est plus important d'extraire dans la couche texte tous les caractères même si certains correspondent à du bruit. Ainsi la table 1 présente notre évaluation en terme de taux de rappel. L'approche est comparée à la méthode [TOM 02] qui propose un post-traitement à l'approche de Fletcher et Kasturi et basée sur des critères géométriques uniquement. Nous pouvons constater l'apport de la représentation parcimonieuse. Le taux moyen de rappel pour notre approche sur l'ensemble de la base d'exemples est de 94% et de 80% pour celle de Tombre et al. L'exemple le plus flagrant et qui montre aussi notre apport est illustré sur l'image 5 de la figure 4 où quasi tous les caractères touchent le graphique. Ceci explique le taux de rappel très faible de la méthode [TOM 02] pour cette image.

Image	**# caractère**	**[TOM 02]**	**Notre méthode**
1	53	49 (92.4%)	53 (100%)
2	78	59 (75.6%)	62 (79.5%)
3	78	68 (87.2%)	75 (96.2%)
4	106	92 (86.8%)	104 (98.1%)
5	21	1 (4.8%)	21 (100%)

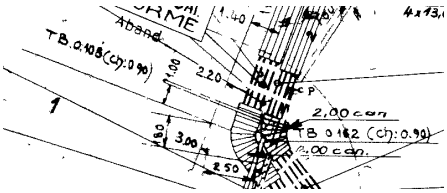
Tableau 1. *Evaluation de performance en terme de taux rappel*



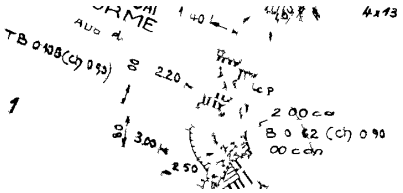
(a) Image 1 : Image originale



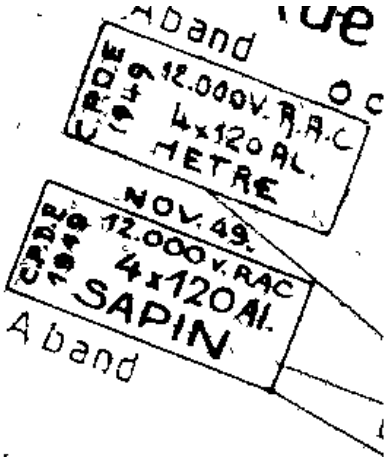
(b) Image 1 : Image binaire texte



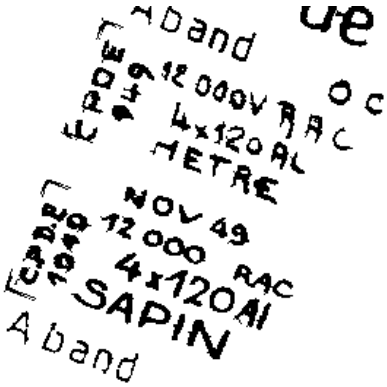
(c) Image 2 : Image originale



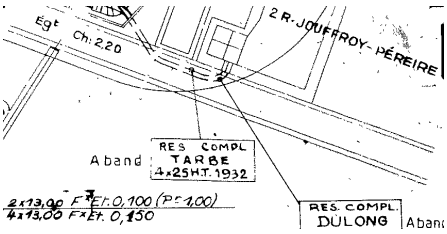
(d) Image 2 : Image binaire texte



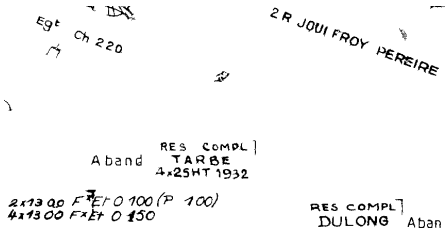
(e) Image 3 : Image originale



(f) Image 3 : Image binaire texte

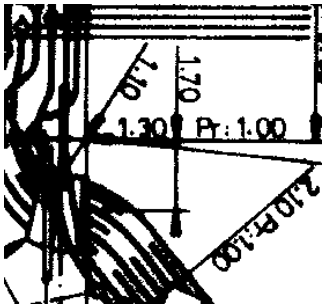


(g) Image 4 : Image originale

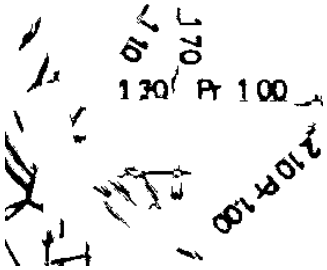


(h) Image 4 : Image binaire texte

Figure 4. *Résultats expérimentaux sur le séparation texte/graphique*



(i) Image 5 : Image originale



(j) Image 5 : Image binaire texte

Figure 4. *Résultats expérimentaux sur le séparation texte/graphique*

Les heuristiques présentées au paragraphe 4 ont été évaluées sur les trois images de la figure 5 (la première colonne). Les chaînes de texte groupés sont données dans les trois images correspondant à la deuxième colonne. La plupart des chaînes de texte contenant des caractères différents et des orientations différentes ont été regroupées avec succès. La seule exception est la chaîne “PTT(0.60)” dans l’image 5d. La raison de cette situation est la connexion entre les caractères ‘6’ et ‘0’ considérés comme un seul caractère car fusionnés et l’orientation de ce caractère est différente des caractères voisins. De plus, le caractère ’)’ est enfouit dans la couche graphique et donc partiellement retrouvé dans la couche texte. Nous pouvons aussi constater que les caractères liés à la ponctuation de la cotation comme le ‘.’ ou ‘:’ont été récupérés dans la chaîne de caractères. Une évaluation quantitative de la méthode de regroupement est donnée dans la table 2.

Image	# **chaînes dans la vérité terrain**	# **chaînes regroupées**
1	15	15 (100%)
2	13	12 (92.3%)
3	9	9

Tableau 2. *Performance de regroupement des caractères extraits*

6. Conclusion

Ce papier propose une méthode qui s’attaque au problème difficile d’extraction de texte dans des documents techniques à partir d’une représentation parcimonieuse. Deux dictionnaires redondant sont utilisés pour capturer les structures texte et graphique du document. L’un est basé sur une transformée d’ondelettes non décimée et l’autre sur la transformée en curvelets. L’algorithme d’analyse en composantes morphologiques (MCA) est utilisé pour optimiser le modèle de décomposition parcimo-

nieux. Les résultats expérimentaux ont montré la robustesse de la méthode qui a donnée sur des exemples complexes un taux de rappel élevé. Nous avons aussi proposé une méthode originale pour regrouper les caractères extraits qui sont alignés en chaîne de caractères.

7. Bibliographie

[AHA 06] AHARON M., ELAD M., BRUCKSTEIN A., « K-SVD : An algorithm for designing overcomplete dictionaries for sparse representation », *IEEE Trans. on Signal Processing*, vol. 54, n° 11, 2006, p. 4311–4322.

[BAR 89] BARLOW H., « Unsupervised learning », *Neural Computation*, vol. 1, n° 3, 1989, p. 295-311.

[CAN 02] CANDÈS E. J., DONOHO D. L., « New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities », *Comm. Pure Appl. Math*, vol. 57, n° 2, 2002, p. 219-266.

[CAO 01] CAO R., TAN C. L., « Text/graphics separation in maps », BLOSTEIN D., KWON Y.-B., Eds., *GREC*, vol. 2390 de *LNCS*, Springer, 2001, p. 167-177.

[CHE 98] CHEN S. S., DONOHO D. L., SAUNDERS M. A., « Atomic decomposition by basis pursuit », *SIAM J. Sci. Comput.*, vol. 20, n° 1, 1998, p. 33-61.

[CHE 04] CHENG Z., CHEN M., LIU Y., « A robust algorithm for image principal curve detection », *Pattern Recognition Letters*, vol. 25, n° 11, 2004, p. 1303-1313.

[DAV 97] DAVIS G., MALLAT S., AVELLANEDA M., « Adaptive greedy approximations », *J. Constr. Approx.*, vol. 13, n° 1, 1997, p. 57–98.

[DEF 94] DEFORGES O., BARBA D., « A robust and multiscale document image segmentation for block line/text line structures extraction », *Proceedings of 12th ICPR*, vol. 2, 1994, p. 306-310.

[DON 94] DONOHO D. L., JOHNSTONE I. M., « Ideal spatial adaptation by wavelet shrinkage », *Biometrika*, vol. 81, n° 3, 1994, p. 425–455.

[DON 06] DONOHO D. L., « For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution », *Comm. Pure Appl. Math*, vol. 59, 2006, p. 797–829.

[ELA 05] ELAD M., STARCK J., QUERRE P., DONOHO D., « Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA) », *Applied and Computational Harmonic Analysis*, vol. 19, 2005, p. 340-358.

[FAD 09] FADILI M., STARCK J.-L., MURTAGH F., « Inpainting and zooming using sparse representations », *Comput. J.*, vol. 52, n° 1, 2009, p. 64–79.

[FLE 88] FLETCHER L. A., KASTURI R., « A robust algorithm for text string separation from mixed text/graphics images », *IEEE Trans. PAMI*, vol. 10, n° 6, 1988, p. 910-918.

[FRE 75] FREEMAN H., SHAPIRA R., « Determining the minimum-area encasing rectangle for an arbitrary closed curve », *Commun. ACM*, vol. 18, n° 7, 1975, p. 409–413.

[GLO 92] GLOGER J., « Use of the Hough transform to separate merged text/graphics in forms », *Proceedings of 11th ICPR*, vol. 1, 1992, p. 268-271.

[GON 01] GONZALEZ R. C., WOODS R. E., *Digital image processing*, Prentice Hall, 2nd édition, 2001.

[HOA 09] HOANG T. V., TABBONE S., PHAM N.-Y., « Extraction of Nom text regions from stele images using area Voronoi diagram », *Proceedings of 10th ICDAR*, 2009, p. 921-925.

[JUT 91] JUTTEN C., HERAULT J., « Blind separation of sources, Part 1 : an adaptive algorithm based on neuromimetic architecture », *Signal Process.*, vol. 24, n° 1, 1991, p. 1–10.

[LI 00] LI L., NAGY G., SAMAL A., SETH S. C., XU Y., « Integrated text and line-art extraction from a topographic map », *IJDAR*, vol. 2, n° 4, 2000, p. 177-185.

[LU 98] LU Z., « Detection of text regions from digital engineering drawings », *IEEE Trans. PAMI*, vol. 20, n° 4, 1998, p. 431-439.

[LUO 97] LUO H., KASTURI R., « Improved directional morphological operations for separation of characters from maps/graphics », TOMBRE K., CHHABRA A. K., Eds., *GREC*, vol. 1389 de *LNCS*, Springer, 1997, p. 35-47.

[OLS 96] OLSHAUSEN B. A., FIELD D. J., « Emergence of simple-cell receptive field properties by learning a sparse code for natural images », *Nature*, vol. 381, 1996, p. 607–609.

[OLS 98] OLSHAUSEN B. A., FIELD D. J., « Sparse coding with an overcomplete basis set : A strategy employed by V1 ? », *Vision Research*, vol. 37, 1998, p. 3311–3325.

[PAN 07] PAN W., BUI T., SUEN C., « Text segmentation from complex background using sparse representations », *Proceedings of 9th ICDAR*, 2007, p. 412-416.

[SAR 00] SARDY S., BRUCE A. G., TSENG P., « Block coordinate relaxation methods for nonparametric wavelet denoising », *J. of Comput. Graph. Stat.*, vol. 9, 2000, p. 361-379.

[SHE 92] SHENSA M., « The discrete wavelet transform : wedding the à trous and Mallat algorithms », *IEEE Trans. on Signal Processing*, vol. 40, n° 10, 1992, p. 2464-2482.

[SON 02] SONG J., SU F., TAI C.-L., CAI S., « An object-oriented progressive-simplification-based vectorization system for engineering drawings : model, algorithm, and performance », *IEEE Trans. PAMI*, vol. 24, n° 8, 2002, p. 1048-1060.

[STA 05] STARCK J.-L., ELAD M., DONOHO D. L., « Image decomposition via the combination of sparse representations and a variational approach », *IEEE Trans. on Image Processing*, vol. 14, n° 10, 2005, p. 1570-1582.

[SU 09] SU F., LU T., YANG R., CAI S., YANG Y., « A character segmentation method for engineering drawings based on holistic and contextual constraints », *Proceedings of 8th GREC*, 2009, p. 280-287.

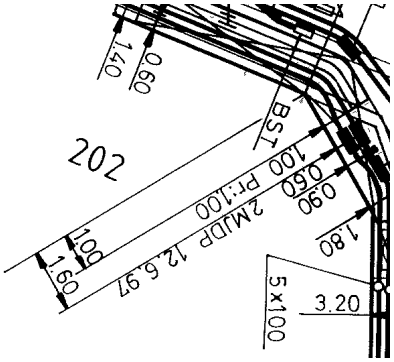
[TAB 06] TABBONE S., WENDLING L., SALMON J.-P., « A new shape descriptor defined on the radon transform », *Comput. Vis. Image Underst.*, vol. 102, n° 1, 2006, p. 42–51.

[TAN 98] TAN C. L., NG P. O., « Text extraction using pyramid », *Pattern Recognition*, vol. 31, n° 1, 1998, p. 63-72.

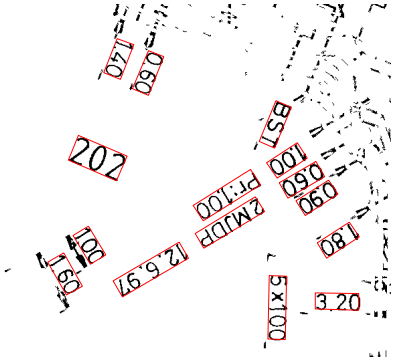
[TOM 02] TOMBRE K., TABBONE S., PÉLISSIER L., LAMIROY B., DOSCH P., « Text/graphics separation revisited », LOPRESTI D. P., HU J., KASHI R. S., Eds., *DAS*, vol. 2423 de *LNCS*, Springer, 2002, p. 200-211.

[VEL 03] VELÁZQUEZ A., LEVACHKINE S., « Text/graphics separation and recognition in raster-scanned color cartographic maps », LLADÓS J., KWON Y.-B., Eds., *GREC*, vol. 3088 de *LNCS*, Springer, 2003, p. 63-74.

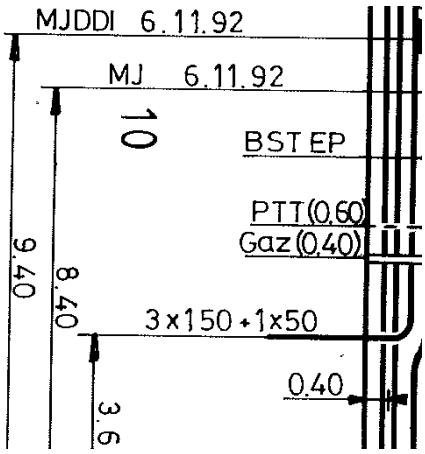
[WAN 01] WANG Y., PHILLIPS I., HARALICK R., « Using area Voronoi tessellation to segment characters connected to graphics », *Proceedings of 4th GREC*, 2001, p. 147–153.



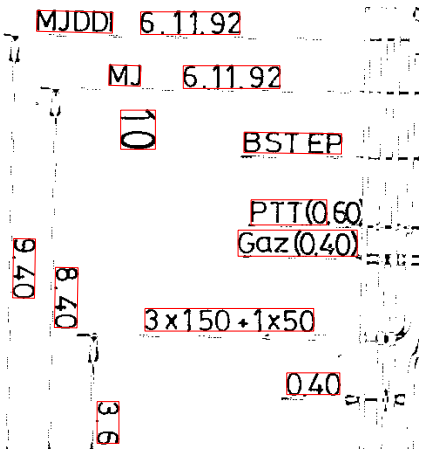
(a) Image 1 : Image originale



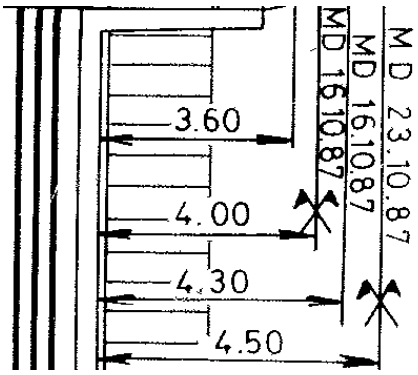
(b) Image 1 : Chaînes de caractères



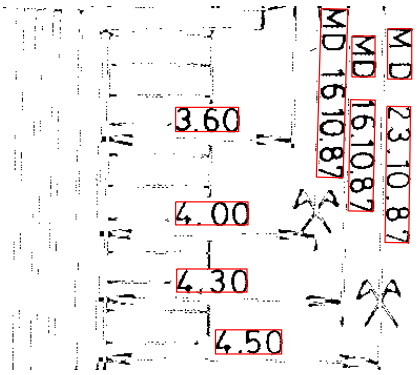
(c) Image 6 : Image originale



(d) Image 6 : Chaînes de caractères



(e) Image 7 : Image originale



(f) Image 7 : Chaînes de caractères

Figure 5. *Résultats expérimentaux sur le regroupement des chaînes de caractères*